贝叶斯因子序列分析：实验设计中平衡信息与效率的新方法

郑元瑞 [1,2] 胡传鹏 [1*]

([1] 南京师范大学心理学院，南京 210024)

([2] 昆明城市学院，昆明 650106)

([*]通讯作者：胡传鹏，E-mail: hcp4715@hotmail.com)

摘 要

实验设计的关键是平衡信息量与效率。贝叶斯因子序列分析利用贝叶斯因子不断更新证据的特点，通过贝叶斯因子标准和在收集数据过程的序列分析来平衡信息量与效率。本文展示如何使用开源软件 JASP 和 R 实现该分析的三个步骤：确定关键效应、确定停止标准、数据收集中序列分析并决策。该方法考虑现实条件且简单易行，可帮助研究者更有效进行实验。

关键词 贝叶斯因子序列分析；统计检验力；实验设计；JASP；R

中图分类号：B849

1. 前言

实验是科学心理学重要方法之一。实验设计的关键是平衡效率和信息量，即如何使用较小的实验成本，如被试量、时间、金钱等的情况下，能够尽可能可靠地检测出感兴趣效应，获得能够拒绝或接受某个研究假设的证据(Stefan et al., 2019)。当实验任务以及测量方式等确定后，对实验的信息量能够产生重大影响的通常是实验中的样本量。由于科学心理学研究中的效应往往比较小(Funder & Ozer, 2019; Richard et al., 2003, 2003; Schäfer & Schwarz, 2019; Thielmann et al., 2020; Götz et al., 2022)，如何选择合适的样本量已成为实验设计中需要研究者重点思考的因素 (温忠麟等, 2022)。

当前，研究者对样本量的估计主要基于零假设显著性检验（null hypothesis significance test , NHST）框架下的先验统计检验力分析（prior power analysis)。该方法有两个明显的不足。首先，其仅关注实验设计的部分信息量，即如何控制假阳性和假阴性的错误率。这导致研究者忽略实验设计的其他方面信息，如支持证据的强度和实验的效率问题等(Schönbrodt et al., 2017; Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019)。其次，该方法过于依赖效应量的设定。心理学可重复危机(胡传鹏等, 2016; Lindsay, 2020; Open Science Collaboration, 2015)，更增加了设定合适的效应量的难度。如果按照已发表的论文进行效应量设定，则可能由于出版偏见等原因(胡传鹏等, 2016)而选择了高估的效应量，导致统计检验力分析的结果脱离真实情况。另一方面，如果选择的效应量过于保守，则可能导致研究者需要承担更多的实验成本(Lakens, 2022)。

近年来，研究者提出贝叶斯因子序列分析(sequential Bayes Factor analysis)以平衡信息量和效率。研究者通过数据模拟发现，在合理地设置先验(prior)和贝叶斯因子(Bayes Factor, BF)决策阈值之后，贝叶斯因子序列分析可以很好地控制实验的假阳性率和假阴性率(Schönbrodt & Wagenmakers, 2018)。由于贝叶斯因子可以监控证据的强度的特性，使用贝叶斯因子序列分析也能为研究者提供丰富的关于当前实验的信息(胡传鹏等, 2018; Schmalz et al., 2021)。

本文将在介绍贝叶斯因子基本概念的基础上，介绍贝叶斯因子序列分析及其应用步骤。随后介绍如何使用开源软件 JASP 和 R 语言对实证数据进行分析。最后，在此基础上探讨贝叶斯因子序列分析的应用前景和不足。

2. 贝叶斯因子序列分析的原理

贝叶斯因子是贝叶斯统计框架下进行假设检验的方法。贝叶斯统计与经典频率主义统计最大的区别在于对概率(probability)的理解不同：在贝叶斯统计框架之下，某个事件的概率

是一个参数，其分布反映了分析者对某个事件的信念强度。分析者对事件的信念也会随着新数据的输入而不断更新；而经典频率主义则认为某事件的概率是一个客观存在的特定值。零假设显著性检验基于特定的统计模型（H0），依据当前的数据模式或更极端的模式出现在这一统计模型之下的可能性而进行推断。相反，贝叶斯因子 BF_{10} 不仅量化了观察数据出现在 H_1 和出现在 H_0 的可能性的比值（见公式 1），而且也反映了当前数据在多大程度上更新了分析者的信念（见公式 2）(Mani et al., 2021; Schönbrodt et al., 2017; Tendeiro, 2022)。

$$BF_{10} = \frac{p(data|H_1)}{p(data|H_0)}$$

公式 1

$$\frac{p(H_1|data)}{p(H_0|data)} = \frac{p(H_1)}{p(H_0)} \times BF_{10}$$

公式 2

由于对 NHST 中 p 值的误解，不少研究者会将 p 值大于或小于显著性水平作为支持实验假设的证据(胡传鹏等, 2016;王珺等, 2021; X.-K. Lyu et al., 2020; Z. Lyu et al., 2018)。这一误解与研究者在数据分析时可能存在的自由度结合（如不断收集数据检查 p 值是否小于 0.05），进而导致荒谬的结论（如认为人类可以预测未来，见 LeBel & Peters, 2011）。为解决这一问题，有研究者提出使用贝叶斯因子假设检验来替代 NHST(吴凡等, 2018; Wagenmakers et al., 2011)。贝叶斯因子的优势在于能同时考虑对零假设和备择假设的支持程度。此外，贝叶斯因子还有监控证据强度的变化、结果会趋于稳定等优点(胡传鹏等, 2018; Wagenmakers et al., 2016)。自 2012 年以来，随着图形界面的软件以及简单易用的 R 工具包(如 JASP, jamovi, BayesFactor)的出现，贝叶斯因子开始被广泛使用。正是由于这些优点和工具，研究者对贝叶斯因子的使用更简便，能动态地了解证据强度的变化，从而实现实验设计中信息量与效率的平衡。

2. 1 贝叶斯因子序列分析

序列分析指的是研究者在数据收集前，根据研究设计选择适当的统计模型。在保证研究获得足够信息的前提下，设置停止数据收集的标准。随后在数据收集过程中，能对新收集的

数据进行持续或阶段地分析。当这些中期分析(interim analysis)的结果达到预定标准时，可及时停止收集数据。由于序列分析在研究中并没有固定的样本量，研究者可以在信息量足够的前提下及时停止收集数据，减少实验成本。

序列分析与 *p* 值操纵中的手段之一——收集数据直到结果显著为止——不同。前者考虑停止规则对信息的影响，将假阳性和假阴性保持在一个可接受的范围(Lakens et al., 2021; Schönbrodt et al., 2017)；而后者则是以传统意义上的“统计显著性”（$p < 0.05$）作为标准，未考虑随着中期分析中检验次数的增加，一类错误出现的风险将大大增加(胡传鹏等, 2016; Ioannidis, 2005; Lakens, 2014; Yu et al., 2014)。在贝叶斯框架和频率主义统计的框架之下，均可以进行序列分析。与频率主义的序列分析中需要提前确定中期分析次数(Lakens, 2022)相比，贝叶斯因子的序列分析操作更为简便。

贝叶斯因子的序列分析能够很好控制假阳性或假阴性，在理论上和实践上均得到了支持。理论上，研究者认为，选择性停止不会影响贝叶斯因子作为量化证据强度的指标 (Rouder, 2014; Schmalz et al., 2021)。通过贝叶斯因子序列分析获得的数据与一次收集数据得到的结果基本一致。这是由于尽管设定的停止阈值会让一些数据（如在 NHST 框架下导致 *p* 值暂时小于 0.05 的数据）更容易被观察到，但在贝叶斯框架下，数据增加会同时改变零假设和备择假设下观察到该数据的可能性，使得两个假设的似然比（也就是贝叶斯因子）不会受到影响(Bayarri et al., 2016)。

实践上，研究者通过模拟对贝叶斯因子序列分析能否能够控制假阳性和假阴性进行了检验。Schönbrodt 等(2017)以独立样本 *t* 检验为例进行了一系列模拟测验。结果表明：当效应量为 0 时，以尺度参数(scale parameter)$r = \sqrt{2}/2$的柯西分布作为先验，将停止收集数据的 BF 阈值设定为 10 或 1/10 时，假阳性率为 4.3%，与传统意义上的 5%的假阳性率相当；当效应量为 0.2 时，以同样的先验分布和 BF 阈值进行贝叶斯因子分析，假阴性率为 5.6%，类似于 NHST 中的统计检验力达到 94.4%。该研究还发现在不同的效应量下，随着柯西分布的先验尺度参数或 BF 阈值增加，假阴性率会逐渐降低为 0，假阳性率会逐渐降低至极小值(Schönbrodt et al., 2017)。同时，在相同效应量前提下，贝叶斯因子序列分析只需传统频率学派的统计检验力分析所需要样本的 30%~50%即可检测出效应的存在，而假阳性或假阴性率与传统统计检验力相同或更低(Schönbrodt et al., 2017)。后续的模拟研究也证明使用贝叶斯因子序列分析并在达到研究者最大可收集样本量时停止继续收集数据的情况下也可以很好地控制假阳性率和假阴性率(Schönbrodt & Wagenmakers, 2018)。

2.2 使用贝叶斯因子序列分析的步骤

贝叶斯因子序列分析大致可分为三个步骤:确定关键的效应、确定停止数据收集的标准、在数据收集中进行分析并决定是否停止收集数据（见图 1）。

贝叶斯因子的序列分析首先需要确定研究中关键的效应。这一点与传统的统计检验力分析相似，但常为研究者所忽视。也就是说，研究者要将最关键的研究假设与统计检验建立联系(Scheel et al., 2021)。

当研究者明确自己需要检验的效应以及对应的统计分析后，下一步是设定 BF 阈值。研究者可以根据先前的模拟研究或自己进行模拟研究来确定这个阈值。通常研究者认为停止收集数据的 BF 阈值为 6 或 1/6 或更严格的 10 或 1/10 时，能够较好地平衡假阴性和假阳性(Moerbeek, 2021; Schönbrodt et al., 2017; Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019)。在设置 BF 阈值时，研究者需要清晰地说明其先验选择，或通过稳健性分析（robustness analysis）得到多个先验下的 BF 值，并将停止收集数据标准设定为多个先验分布下的 BF 值均达到阈值，从而更加有把握地进行推断。

研究者也需要考虑现实条件对数据收集的制约。Schönbrodt 和 Wagenmakers(2018)建议，研究者可以根据现实条件考虑自己在实验中能够收集的最大样本作为停止收集数据的另一个标准。同时，为避免使用贝叶斯因子序列分析在收集数据早期停止时出现估计效应量最大条件偏差以及较大的假阳性，可以先收集每组 12 或 20 个的最小样本量来避免误导性的证据(Schönbrodt et al., 2017; Svensson et al., 2021)。

第三步，理论上讲，一旦达到先前设定的 H_1 或 H_0 的阈值，研究者应该停止收集数据并且报告最终的 BF_{10}，也可报告均值和设定效应量的实际等价区以及最高后验密度区间(HDI)(许岳培等, 2021)，或对整个后验分布作图(Schönbrodt et al., 2017)。实际研究中，当达到其他实验前确定的标准时也可以停止，如达到最大样本量。此外，当 BF 达到先前设定阈值时，可继续让已经预约的被试完成实验。

以上三个步骤为开放式贝叶斯因子序列分析的主要步骤。研究者还可在事先定义停止阈值和先验效应量分布前进行预注册(Svensson et al., 2021)。

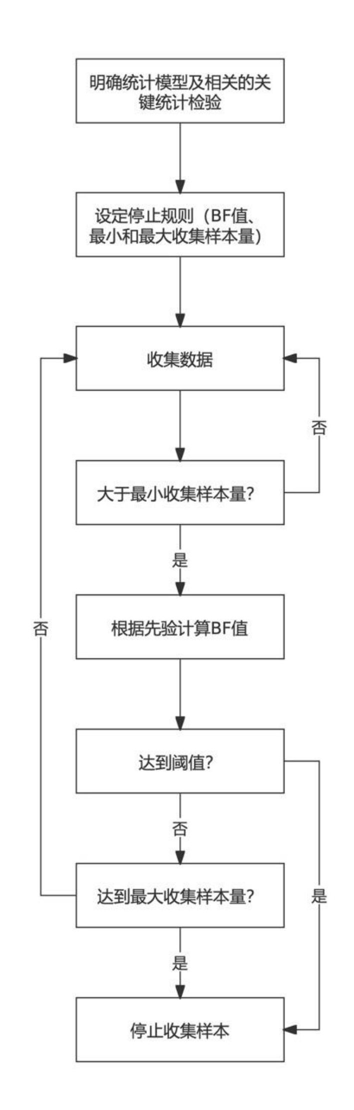


图 1: 贝叶斯因子序列分析流程图

3. 贝叶斯因子序列分析的实现

3.1 示例数据

为演示贝叶斯因子序列分析，本文使用一组通过 jsPsych (de Leeuw, 2015) 在线收集的数据。该在线实验采用社会联结范式(Sui et al., 2012)。实验中，被试首先学习三个图形（三角形、圆形和正方形）和三个不同道德效价的人物标签（“好人”、“坏人”和“常人”）之间的联系。随后，被试需要完成一个简单的知觉匹配任务：屏幕上同时呈现的图形和文字标签 100 毫秒，被试需判断呈现的图形与标签是否与属于先前学习到的联结关系，如果属于，则反应为匹配，否则是不匹配。图形与道德标签之间的关系在被试间进行平衡。数据经过预处理后，整理为认知心理学实验中常见的形式，即每个被试的数据仅包括其在每个条件下的反应时均值和正确率。数据可以通过链接：https://github.com/Chuan-Peng-Lab/SBFA_Tutorial 获取。

3.2 步骤 1:确定关键的 BF 值、相应的统计模型以及停止收集数据的标准

本示例中，研究目的是检验不同道德效价水平的人物标签与图形进行联结后，是否会影响被试对随后的知觉匹配任务的知觉决策过程。根据先前研究(Hu et al., 2020; Sui et al., 2012)，不同效价的道德标签在联结学习后对知觉匹配的影响主要体现在匹配的试次中。因此，我们的假设会出现匹配与道德标签交互作用。此外，假定我们对道德标签的效应有非常具体的假设，即积极效价优于中性条件，中性条件优于消极条件。由于知觉匹配任务中既有反应时间（reaction times，RT），也有正确率数据，还有基于信号检测论所估计出的敏感性 *d'*、决策标准 *c* 等参数值。因此，感兴趣的研究假设可能会对应着多个统计检验，研究者需要从中选定最关键的统计检验。在本示例中， 根据先前研究的结果，我们推断反应时间可能是最敏感的指标，因此，在众多的因变量中**选择反应时间的结果**作为贝叶斯因子序列分析中的关键效应。

表 1. 示例数据集中的研究问题、统计假设与关键的 BF 值

研究问题	统计假设	因变量	可能的 BF 值（加粗为关键）
道德效价是否对知觉匹配中的匹配试次产生影响	匹配与效价的交互作用	**RT**	**2×3 RM ANOVA 中的交互作用**
	匹配与效价的交互作用	ACC	2×3 RM ANOVA 中的交互作用
	匹配与效价的交互作用	*d* prime	One-way RM ANOVA
	积极匹配条件优于中性匹配条件	**RT**	***t*-test**
	积极匹配条件优于中性匹配条件	ACC	*t*-test
	积极匹配条件优于中性匹配条件	*d* prime	*t*-test
	消极匹配条件劣于中性匹配条件	**RT**	***t*-test**
	消极匹配条件劣于中性匹配条件	ACC	*t*-test
	消极匹配条件劣于中性匹配条件	*d* prime	*t*-test

* 在一个简单的 2×3 被试内实验设计中，考查一个自变量的效应时，潜在的统计检验可以多达 9 个。本示例中通过先前研究选定关键的研究假设，而非收集数据后探索显著的效应，避免了研究者自由度过大带来

的假阳性(Simmons et al., 2011)。具体而言，我们将关注贝叶斯重复测量方差分析中的交互作用和两个以反应时间作为因变量的配对样本 *t* 检验（表 1 中加粗的部分）。

根据先前的模拟研究，我们确定停止继续收集样本的 BF 阈值：BF_{10} 阈值为 10 或 1/10。此时，有足够的证据支持备择假设或原假设，可停止收集数据。

同时，我们也根据 Schönbrodt 等(2017)的建议和现实收集数据的制约因素，假定本次研究中最小收集样本量和最大收集样本量分别设定为，12 名被试和 20 名被试[1]。

3.3 步骤 2：中期分析

3.3.1 *t* 检验

首先，使用 JASP 软件对数据进行配对样本 *t* 检验的贝叶斯因子序列分析分析。读取数据等步骤可以参照胡传鹏等(2018)的教程。以下将以示例数据为例，使用 JASP 0.16.4 进行展示(Love et al., 2019)。由于关键效应有两个单侧 *t* 检验，因此，在中期分析中， 我们需要对两个配对样本 *t* 检验的结果进行持续分析。

然后，为在 H_1 条件下的效应量选择一个先验分布，在 JASP 和 R 语言 BayesFactor 工具包中，默认使用默认先验分布(default priors)。之后通过点击界面上的 T-Tests 面板，选择“Bayesian”分类下的“Paired samples T-Tests”，将变量“RT_Good_Match”和“RT_Neutral_Match”放入到“Variable Pairs”框内，根据前文的假设，选择“Alt.Hpothesis”下的“Measure 1 < Measure 2”进行单侧配对样本 T 检验，并点击“Plots”下方的“Sequential analysis”以及子选项“Robustness check”对数据进行贝叶斯因子序列分析和稳健性检验，步骤和结果见图 2 A。

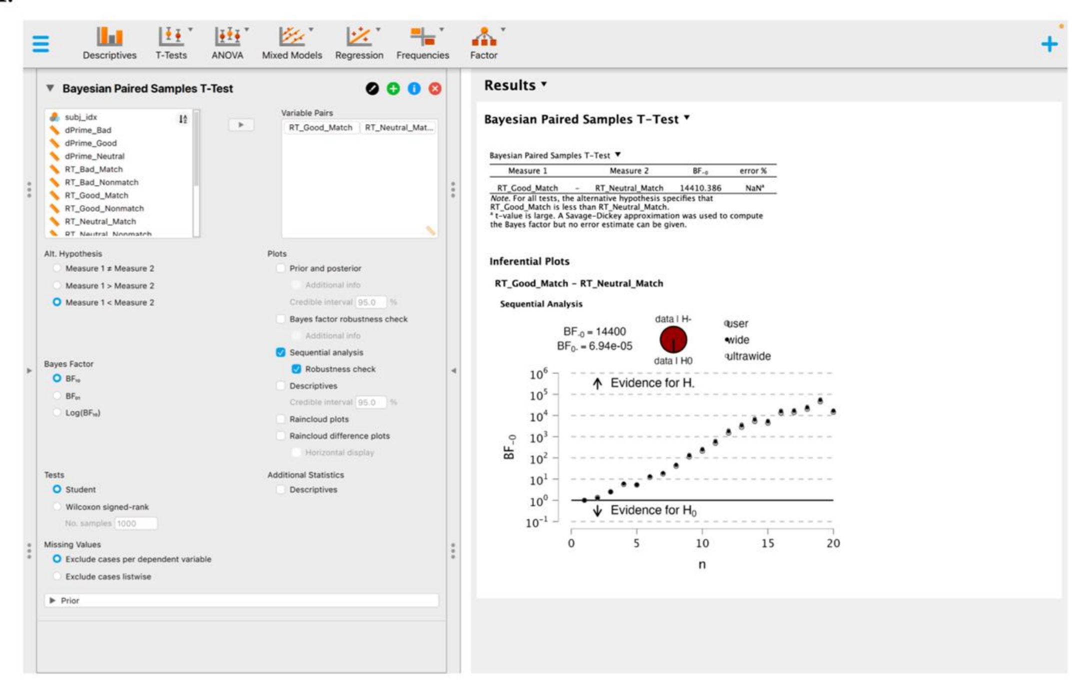
如图 2 A 右侧结果显示，在收集到第 6 个被试时，积极匹配条件和中性匹配条件比较的单侧配对样本 *t* 检验的 BF_{10} 值超过 10，表明有极强的证据支持 H_1。

同时，对另一个假设“消极匹配条件劣于中性匹配条件”进行 BF 配对样本 *t* 检验，仅需选择对应的变量并改变在“Alt.Hpothesis”下选择“Measure 1 > Measure 2”的选项就可以得到与假设符合的贝叶斯因子序列分析的结果。如图 2 B 右侧结果显示，在收集样本达到最大收集样本量 20 名被试时，虽然消极匹配条件和中性匹配条件比较的单侧配对样本 *t* 检验的 BF_{10} 值有明显偏向 H_0 的阈值，但是并没有达到或超过事先定义的 H_0 阈值 1/10。同时敏感性分析表明，在先验分布较为宽时，BF_{10} 接近 1/10。这说明，研究者需要评估现实因素，判断是否

[1] 此处最大样本量的设置是出于演示的目的，12 与 20 对真实实验不具备参考价值。研究者应根据实际条件来设定最大样本量，比如采集数据所需要的时间、财力、物力和人力。例如，单个实验中收集 50 名被试的有效脑电数据可能就是实际上能够接受的最大样本量。

可以接受继续收集数据直到 BF_{10} 达到设定的阈值。

A.



B.

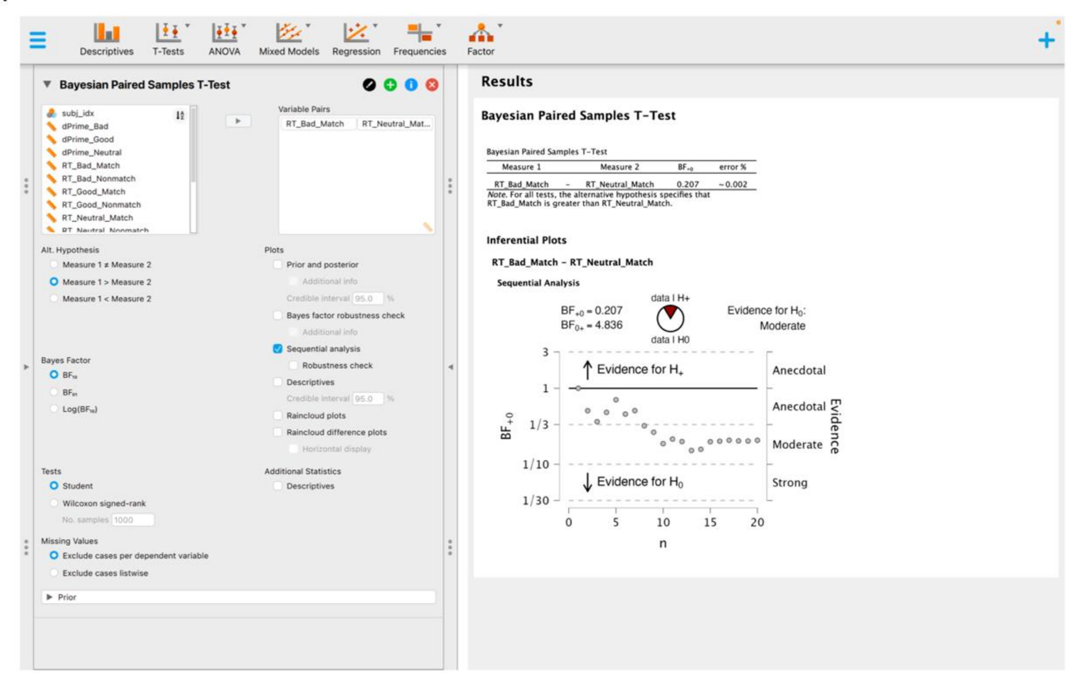


图 2: 使用 JASP 进行配对样本 t 检验的截图。（A）对积极匹配条件优于中性匹配条件这一假设进行 t 检验序列分析的操作截图; (B) 对消极匹配条件劣于中性匹配条件这一假设进行贝叶斯因子 t 检验序列分析的操作截图。

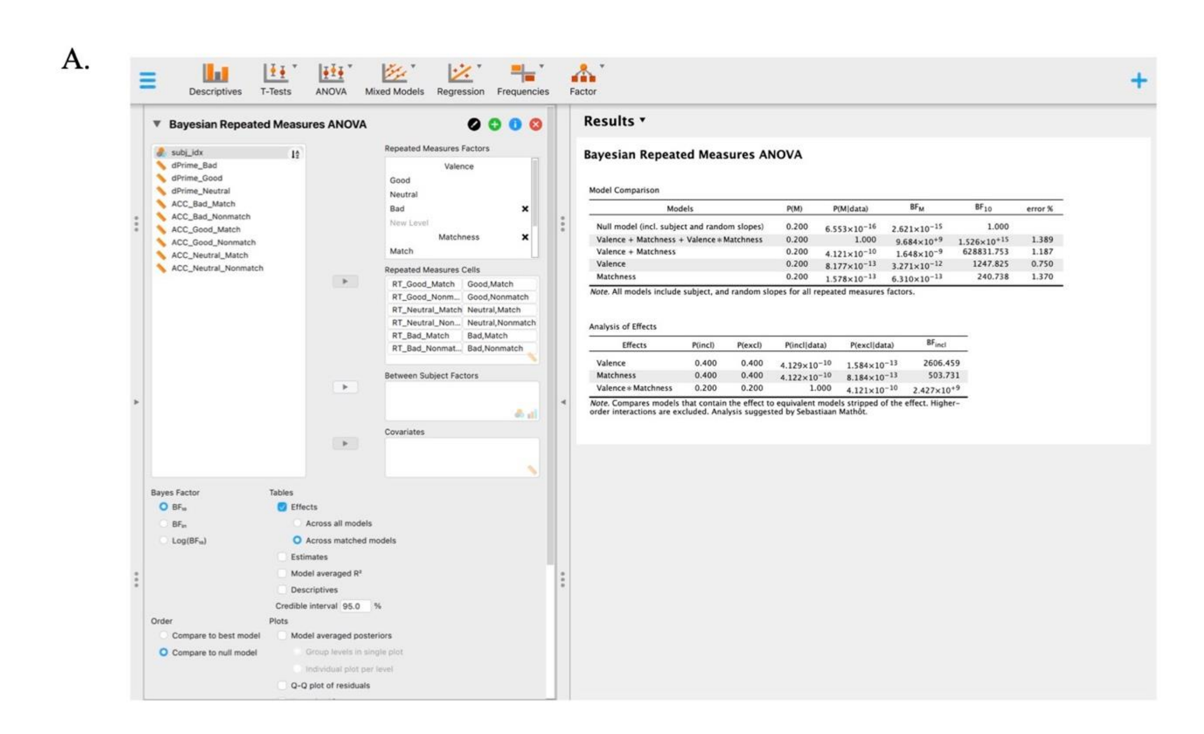
3.3.2 方差分析

在本研究中，检验匹配与道德标签是否存在交互作用也是关键效应，因此，在数据收集过程中我们也将持续地进行贝叶斯因子重复测量方差分析并监测其结果。

贝叶斯（重复测量）方差分析是用于比较模型包含或不包含某一因子和交互项的预测能力表现(Keysers et al., 2020)。在进行贝叶斯重复测量方差分析时，要先点击 JASP 界面上的 ANOVA 面板，选择“Bayesian”分类下的“Repeated Measures ANOVA”，之后在“Repeated Measures Factors”框中，将被试内变量的名称进行命名，命名结束后，在该框下方“Repeated Measures Cells”中会出现命名后的条件，将数据中的变量放入对应的条件框中即可，如图 3A 左侧所示（其他类型的贝叶斯 ANOVA 分析，见王允宏等, 2022）。

图 3A 右侧最上方的表格表示使用贝叶斯因子来比较所有的模型与最简单的零模型进行比较。在本例中，计算交互项的效应需要比较包含交互作用的模型，即 Valence+Matchness+Valence×Matchness 和不包含交互作用的模型，即 Valence+Matchness，得到 BF_{10} = 2426722239.64，有极强的证据支持交互项的效应。在本示例中，由于交互作用非常明显，仅收集 3 个被试的数据后，交互作用达到了先前设定的阈值。如果我们停止数据收集的标准仅关注交互作用，则可以停止收集数据。但由于本研究先前设定的停止标准是 3 个检验的 BF 均达到阈值或达到最大样本量，因此继续收集数据并监测重复测量 ANOVA 的 BF 值。值得注意的是，JASP 未提供重复测量方差分析序列分析的可视化效果。由于 JASP 使用 R 包 BayesFactor 作为底层的工具(Wagenmakers et al., 2018)，我们也可以使用 R 代码完成对重复测量方差分析中交互作用的 BF_{10} 值变化趋势的可视化（见图 3）。

A.



B.

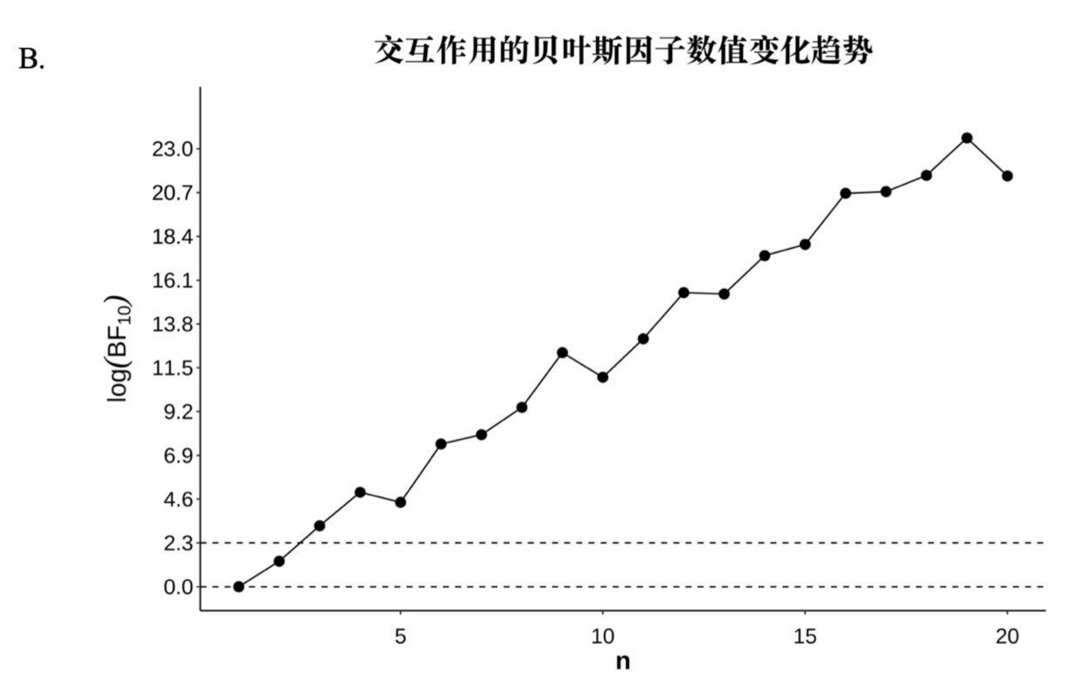


图 3 使用重复测量方差分析来检验匹配与效价的交互作用。（A）JASP 的操作截图；（B）交互作用的 BF_{10} 值随着样本增加的而变化，y 轴刻度是以常数 e 为底数进行 log 变换后的 BF_{10} *的数值*，结果保留一位小数。代码见：https://github.com/Chuan-Peng-Lab/SBFA_Tutorial

综合以上贝叶斯因子序列分析的结果，由于关键统计假设之一"消极匹配条件劣于中性匹配条"统计结果的 BF_{10} 值并未达到事先设定的阈值，理论上讲研究还需要继续收集数据，

直到三个关键统计假设效应的 BF 值均达到事先设定的阈值为止。但本演示数据中同时规定最大样本量为 20，因此最终样本量为 20。

3.4 步骤 4：论文中的报告

为了演示完整报告贝叶斯因子序列分析结果，根据 van Doorn 等(2021)建议在方法部分关于样本的信息，进行如下描述：

"*使用贝叶斯因子序列分析方法确定本实验的样本量(Schönbrodt et al., 2017)，停止收集数据的贝叶斯因子阈值 BF_{10} 设定为 10（接受 H_1 的阈值）和 1/10（接受 H_0 的阈值）。具体而言，为了检验"道德效价是否对知觉匹配中的匹配试次产生影响"这一假设，本研究关注不同条件下反应时间的差异，采用贝叶斯因子序列设计检验三个关键效应所对应的 BF 值：道德效价与匹配程度的交互作用、"道德积极的匹配条件的反应时间快于中性匹配条件"和"道德消极的匹配条件下反应时间慢于中性的匹配条件"。这三个关键效应将分别采用贝叶斯因子重复测量方差分析和贝叶斯配对样本 t 检验（单侧）。贝叶斯因子重复测量方差分析和贝叶斯配对样本 t 检验均使用 JASP 内置先验(Rouder et al., 2009)。考虑到本研究的现实因素，将收集的最小样本量设定为 12 名被试，最大可收集的样本量设定为 20 名被试。*

数据收集过程中，当收集样本达到最小样本量 12 名被试时进行第一次分析，如果所有 BF_{10} 值超过两个阈值之一时停止收集数据。达到最大收集样本量 20 名被试也停止收集数据，三个关键 BF 中两个已经大于设定的停止标准，另一个虽然没有大于设定的停止标准，但 BF 值显示有较强证据支持 H_0，即消极匹配条件与中性匹配条件在反应时间上没有差异。"

需要注意的是在停止收集数据规则需要研究者根据研究问题进行在数据收集前进行设定（如在预注册中进行设定）。其中，与研究假设对应的关键 BF 值的选择、统计模型（重复测量方差分析、*t* 检验还是回归等）以及 BF 阈值选择尤其需要进行说明。最大样本量和最小样本量也需要根据实际情况进行合理解释。

使用贝叶斯因子序列分析进行样本量规划时，某种程度上已经完成了关键效应进行评估。但通常研究者进行实验时并非仅仅对某个单一的效应感兴趣，因此还可以继续进行后续的统计分析，对其他效应（如方差分析中的主效应）进行检验，或对关键效应进一步分析（如对方差分析中的交互项进行简单效应分析）。

4. 总结与展望

实验设计很大程度上决定一个实验研究的质量，而实验的信息量与效率是实验设计中的重要考量。虽然传统基于 NHST 的统计检验力分析试图为研究者提供这样的一个统计工

具，但是这一方法自提出起(Cohen, 1988)并未得到足够的重视，直到可重复性危机后期刊才逐渐开始将样本量的合理说明作为文章方法中必须报告的部分(Simmons et al., 2012)。可能的原因在于进行统计检验力分析本身较为困难。最近一项元研究发现绝大部分发表论文中报告的统计检验力分析的过程难以被重复(McKay et al., 2022)。本文介绍的贝叶斯因子序列分析及其实施步骤：确定关键效应、确定停止标准、数据收集中序列分析并决策，在操作上更加简易且考虑到现实状况。

当然，贝叶斯因子序列分析也并非万能，盲目使用贝叶斯因子来监测证据强度可能会带来问题。首先，虽然在心理学研究中，研究者通常使用两个假设对立的方式来进行检验，但真实生成数据的模型可能是两个或多个假设的混合。加之在使用 ANOVA 对数据进行分析时，备择假设通常是条件之间存在差异，而没有更清晰的假设。在这种情况下，盲目使用贝叶斯因子来监控证据强度可能会增加实验者在他们想要的方向上找到证据的可能性(Sanborn & Hills, 2014)。其次，贝叶斯因子 *t*-test 和 ANOVA 与传统 *t*-test 和 ANOVA 一样，需要满足一定前提预设才能应用。如果忽略前提预设，数据本身的生成模型与贝叶斯因子分析中的统计模型不匹配时，会导致贝叶斯因子分析无法提供有效的方式来控制假阳性(de Heide & Grünwald, 2021; Yu et al., 2014)。当然，这一问题是否为贝叶斯因子序列分析的不足也存在争议(见 Rouder & Haaf, 2019)。最后，虽然使用 JASP 和 R 程序包 BayesFactor 进行贝叶斯因子序列分析较为简便，但当效应量较小，所需样本量较大时，计算所需要的迭代次数也随之增加，需要较高的计算成本(Fu et al., 2021)。

尽管贝叶斯因子序列分析目前没有像先验检验力分析那样在设计实验和规划样本量中被广泛使用，但其能够很好的平衡实验中的信息量和效率的优势，使得越来越多的研究者在规划样本量时使用该方法。贝叶斯因子序列分析已经被运用于婴儿早期词汇学习(Mani et al., 2021)和社会知觉决策(C.-P. Hu et al., 2020) 等研究的样本量规划之中。贝叶斯序列分析能够及时停止数据收集的特点，在资源有限时尤其重要。例如，心理科学加速器（Psychological Science Accelerator, PSA）的第六个项目中，作者团队最初设计通过 PSA 的合作者网络来检验一个关键的研究假设，即道德判断的跨文化差异 (Bago et al., 2022)。但当研究方案作为第一阶段的注册报告提交给 *Nature Human Behavior* 后，审稿人要求额外检验假设。为满足这个要求同时避免给全球合作者带来过大的工作量，方法团队最终使用贝叶斯因子序列分析进行样本量规划。当第一个研究的数据达到了数据停止收集的 BF 阈值（BF_{10}>10 或 BF_{10}<1/10）后，被试被自动引导到第二个实验的链接，继续完成第二个研究数据收集。

贝叶斯因子序列分析可以扩展到更复杂的统计模型，具有广阔的应用前景。例如，当研

究者需要使用层级模型（或线性混合模型）时，如果研究者清楚其研究所关注的效应，也可以通过与 ANOVA 类似的模型比较思路，使用其他工具包（如 brms(Bürkner, 2017)）进行贝叶斯因子序列分析 (Vasishth et al., 2022)。

参考文献

胡传鹏，孔祥祯，Wagenmakers, E. J., Alexander, L. Y.，彭凯平. (2018). 贝叶斯因子及其在 JASP 中的实现. *心理科学进展*, *26*(6), 951-965.

胡传鹏，王非，过继成思，宋梦迪，隋洁，彭凯平. (2016). 心理学研究中的可重复性问题：从危机到契机. *心理科学进展*, *24*(9), 1504-1518.

吴凡，顾全，施状华，高在峰，沈模卫. (2018). 跳出传统假设检验方法的陷阱——贝叶斯因子在心理学研究领域的应用. *应用心理学*, 24(3), 195-202.

王珺，宋琼雅，许岳培，贾彬彬，陆春雷，陈曦，戴紫旭，黄之玥，李振江，林景希，罗婉莹，施赛男，张莹莹，臧玉峰，左西年，胡传鹏. (2021). 解读不显著结果：基于 500 个实证研究的量化分析. *心理科学进展*, 29(3), 381-393.

王允宏, van den Bergh, Don, Aust, Frederik, Ly, Alexander, Wagenmakers, Eric-Jan, 胡传鹏. (2022). 贝叶斯方差分析在 JASP 中的实现. ChinaXiv. https://chinaxiv.org/abs/202209.00140V1.

温忠麟，谢晋艳，方杰，王一帆. (2022). 新世纪 20 年国内假设检验及其关联问题的方法学研究. *心理科学进展*, *30*(8), 1667-1681.

许岳培，陆春雷，王珺，宋琼雅，贾彬彬，胡传鹏. (2021). 评估零效应的三种统计方法. *应用心理学*, *28*(4), 369–384.

Bago, B., Kovacs, M., Protzko, J., Kekecs, Z., Nagy, T., Hoekstra, R., … Aczel, B. (2022). Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nature Human Behaviour*, *6*, 880–895. https://doi.org/10.1038/s41562-022-01319-5

Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, *72*, 90–103. https://doi.org/10.1016/j.jmp.2015.12.007

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

de Heide, R., & Grünwald, P. D. (2021). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review*, *28*(3), 795–812. https://doi.org/10.3758/s13423-020-01803-x

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the Bayesian t test and Welch’s test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods*, *53*(1), 139–152. https://doi.org/10.3758/s13428-020-01408-1

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and

Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small Effects: The Indispensable Foundation for a Cumulative Psychological Science. *Perspectives on Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Hu, C.-P., Lan, Y., Macrae, C. N., & Sui, J. (2020). Good Me Bad Me: Prioritization of the Good-Self During Perceptual Decision-Making. *Collabra: Psychology*, *6*(1), 20. https://doi.org/10.1525/collabra.301

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*(7), 788–799. https://doi.org/10.1038/s41593-020-0660-4

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses: Sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., Pahlke, F., & Wassmer, G. (2021). Group Sequential Designs: A Tutorial. https://doi.org/10.31234/osf.io/x4azm

LeBel, E. P., & Peters, K. R. (2011). Fearing the Future of Empirical Psychology: Bem's (2011) Evidence of Psi as a Case Study of Deficiencies in Modal Research Practice. *Review of General Psychology*, *15*(4), 371–379. https://doi.org/10.1037/a0025172

Lindsay, D. S. (2020). Seven steps toward transparency and replicability in psychological science. *Canadian Psychology / Psychologie Canadienne*, *61*(4), 310–317. https://doi.org/10.1037/cap0000222

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Smíra, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2019). **JASP**: Graphical Statistical Software for Common Statistical Designs. *Journal of Statistical Software*, *88*(2). https://doi.org/10.18637/jss.v088.i02

Lyu, X., Xu, Y., Zhao, X., Zuo, X., & Hu, C. (2020). Beyond psychology: Prevalence of p value and confidence interval misinterpretation across different fields. *Journal of Pacific Rim Psychology*, *14*. https://doi.org/10.1017/prp.2019.28

Lyu, Z., Peng, K., & Hu, C.-P. (2018). P-Value, Confidence Intervals, and Statistical Inference: A New Dataset of Misinterpretation. *Frontiers in Psychology*, *9*, 868. https://doi.org/10.3389/fpsyg.2018.00868

Mani, N., Schreiner, M. S., Brase, J., Köhler, K., Strassen, K., Postin, D., & Schultze, T. (2021). Sequential Bayes Factor designs in developmental research: Studies on early word learning. *Developmental Science*, *24*, e13097. https://doi.org/10.1111/desc.13097

McKay, B., Bacelar, M., & Carter, M. (2022). On the reproducibility of power analyses in motor behavior research. https://doi.org/10.51224/SRXIV.184

Moerbeek, M. (2021). Bayesian updating: Increasing sample size during the course of a study. *BMC Medical Research Methodology*, *21*, 137. https://doi.org/10.1186/s12874-021-01334-6

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science.

Science, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., & Haaf, J. M. (2019). Optional Stopping and the Interpretation of The Bayes Factor. https://doi.org/10.31234/osf.io/m6dhw

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*(2), 283–300. https://doi.org/10.3758/s13423-013-0518-9

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*, 813. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00813

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2021). What is a Bayes factor? *Psychological Methods*. https://doi.org/10.1037/met0000421

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schönbrodt, F. D., Zehetleitner, M., Wagenmakers, E.-J., & Perugini, M. (2017). Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 Word Solution. *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology*, 26, 4-7.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Sui, J., He, X., & Humphreys, G. (2012). Perceptual Effects of Social Salience: Evidence From Self-Prioritization Effects on Perceptual Matching. *Journal of Experimental Psychology. Human Perception and Performance*, *38*, 1105–1117. https://doi.org/10.1037/a0029792

Svensson, J. E., Schain, M., Knudsen, G. M., Ogden, R. T., & Plavén-Sigray, P. (2021). Early stopping in clinical PET studies: How to reduce expense and exposure. *Journal of Cerebral Blood Flow & Metabolism*, *41*(11), 2805–2819. https://doi.org/10.1177/0271678X211017796

Tendeiro, J.N., Kiers, H.A.L. & van Ravenzwaaij, D. (2022). Worked-out examples of the adequacy of Bayesian optional stopping. *Psychon Bulletin & Review* 29, 70–87.

https://doi.org/10.3758/s13423-021-01962-5.
Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), 30–90. https://doi.org/10.1037/bul0000217
van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5
Vasishth, S., Yadav, H., Schad, D. J., & Nicenboim, B. (2022). Sample Size Determination for Bayesian Hierarchical Models Commonly Used in Psycholinguistics. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-021-00125-y
Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7
Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian Benefits for the Pragmatic Researcher. *Current Directions in Psychological Science*, *25*(3), 169–176. https://doi.org/10.1177/0963721416643289
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790
Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*(2), 268–282. https://doi.org/10.3758/s13423-013-0495-z

Sequential Bayes Factor Analysis: Balance Informativeness and Efficiency in Designing Experiments

ZHENG Yuan-Rui[1,2]   HU Chuan-Peng[1]

([1]School of Psychology, Nanjing Normal University, Nanjing, 210024, China)

([2]Department of Applied Psychology, School of Education, Kunming City College, Kunming, 650106, China)

Abstract

The key of experimental design is to balance between informativeness and efficiency. However, power analysis only focuses on informativeness and is difficult to implement. Sequential Bayes Factor analysis takes the advantage of Bayes Factor's ability to continuously update the evidence and reach a trade-off between informativeness and efficiency by setting Bayes Factor criteria and the sequential analysis during data collection. The present primer demonstrates how to perform three steps of sequential Bayes Factor analysis using open-source software JASP and R. This method considers practical issues in real research practices and is easy to implement, which can help researchers to design more efficient experiments.

Key Words: Sequential Bayes Factor Analysis; Power Analysis; Experimental Design; JASP; R